PAPER Special Section on Security, Privacy, Anonymity and Trust in Cyberspace Computing and Communications

Anonymization Technique Based on SGD Matrix Factorization

Tomoaki MIMOTO[†a)], **Seira HIDANO**[†], ***Nonmembers***, **Shinsaku KIYOMOTO**[†], ***and*** **Atsuko MIYAJI**[††], ***Members***

SUMMARY Time-sequence data is high dimensional and contains a lot of information, which can be utilized in various fields, such as insurance, finance, and advertising. Personal data including time-sequence data is converted to anonymized datasets, which need to strike a balance between both privacy and utility. In this paper, we consider low-rank matrix factorization as one of anonymization methods and evaluate its efficiency. We convert time-sequence datasets to matrices and evaluate both privacy and utility. The record IDs in time-sequence data are changed at regular intervals to reduce re-identification risk. However, since individuals tend to behave in a similar fashion over periods of time, there remains a risk of record linkage even if record IDs are different. Hence, we evaluate the re-identification and linkage risks as privacy risks of time-sequence data. Our experimental results show that matrix factorization is a viable anonymization method and it can achieve better utility than existing anonymization methods.

key words: *time-sequence data, anonymization, matrix factorization, privacy and utility*

1. Introduction

Personal data is essential to build an efficient and sustainable society, but at the same time, it is sensitive and must be handled carefully. Time-sequence data, such as purchase and movement history, has attractive values at a macroscopic level. For example, vehicle trajectories are useful for finding the cause of a traffic jam and purchase histories help to create a marketing strategy. In contrast with security, a key challenge of preserving privacy in personal data is that an attacker can be an authorized user, who is the anonymized data receiver, so it is important to keep a balance between privacy and utility and some techniques for achieving this balance are being studied [1].

Time-sequence data includes static and dynamic attributes. Static attibutes are identifiers and quasi-identifiers, such as name, age, and gender. Dynamic attributes are characteristic in time-sequence datasets and they include information concerning time-sequence. A pseudonymized ID, time stamp, location, the direction of movement of a vehicle, engine speed, and number of purchases are included in dynamic attributes as examples.

Manuscript received March 10, 2019.
Manuscript revised October 2, 2019.
Manuscript publicized November 25, 2019.

[†]The authors are with KDDI Research, Inc., Fujimino-shi, 356–8502 Japan.
[††]The author is with Osaka University, Suita-shi, 565–0871 Japan.
a) E-mail: to-mimoto@kddi-research.jp

DOI: 10.1587/transinf.2019INP0013

Dynamic attributes are especially important and has high utility value, but it also presents a high risk leakage of private information even if some information is generalized or deleted. Some previous research such as [2] shows that the amount of processing needed is surprisingly high to preserve the privacy of a time-sequence data.

Furthermore, time-sequence data has a potential risk against an authorized user who is an anonymized data receiver. Time-sequence data may include the same user's data at different time frames. In this case, the primary keys are changed at regular intervals and the records of a user are shuffled with those of other users in many cases. However, there is a tendency for people to take similar action and the privacy risk to link the same users in anonymized datasets. There are few studies concerning about the problem, so we define the attack model and evaluate the risk against an authorized user.

Some time-sequence data such as location data can be denoted as a matrix and we also consider a matrix data. There is some research that manages a time-sequence data as a matrix [1], [3]–[5] and our main proposal is to maintain the utility and the privacy of a dataset using matrix operations.

1.1 Our Contribution

There are two main contributions in this paper. The first one is that we regard a matrix factorization technique as an anonymization method and evaluate the effect to an actual data. Furthermore, we propose an anonymization algorithm which is combined a matrix factorization technique and other anonymization techniques. We apply the algorithm to an actual data and evaluate the privacy risk and the utility. One of the strengths of our proposed algorithm is that it can cahnge the rank r to modulate the privacy risk flexibly compared to previous anonymization algorithms. The other contribution is that we define linkage attack, which is a privacy risk peculiar to a time-sequence data, and evaluate the risk in an actual time-sequence data. The following are the details.

1.1.1 Anonymization Using Matrix Factorization

Matrix factorization is a fundamental step in data analysis. In particular, matrix $M \in \mathbb{R}^{n\times m}$ is decomposed into $U \in \mathbb{R}^{r\times n}$ and $V \in \mathbb{R}^{r\times m}$. U and V represent properties of rows and

columns respectively. The matrix $X = U^{\top}V$ is an approximation of M and rank r affects the accuracy. We observe in our later evaluations that low-rank matrix factorization helps with anonymization, i.e., a low-rank matrix is more likely to withstand re-identification and linkage attacks. We propose to anonymize only U, which is the feature matrix of the users, to maintain the utility because V, which is the feature matrix of the items, does not have any private information. We evaluate the effect of our proposal method in an actural time-sequence dataset in Sect. 5.

1.1.2 Privacy Definition against an Authorized User

Most of the existing research [6] considers that the privacy leaks when an anonymized record is linked to the original record. However, especially in time-sequence data, there is a possibility that an authorized user be an attacker. More precisely, continuity of time-sequence datasets is an important factor in some use-cases, such as medical care, but it may lead to increased risk of re-identification. Therefore, the primary keys are changed at regular intervals and the records of a user are shuffled with those of other users in many cases. However, there is a tendency for people to take similar action. For example, people follow the same trajectory at the same time (e.g., during the commute to work) and there are some people who buy their favorite products, which remain the same over time. Hence, there are some linkage risks between records that represent the same user even if the primary keys are different. Linking the records themselves may lead to leakage of some other information. We define linkage attack in Sect. 3 to address the privacy and evaluate the effect in Sect. 5.

This paper is expanded from [7]. The matrix factorization algorithm is changed and stochastic gradient descent (SGD) based matrix factorization, which is widely used in many fields such as recommender system, is applied.

2. Related Work

The related work presented below is grouped under k-anonymization and noise addition as anonymization methods and we present some matrix factorization techniques applied to time-sequence data.

2.1 Existing Anonymization Methods

2.1.1 k-Anonymization

k-anonymity [8]–[10] is a well-known privacy model. The property of k-anonymity is that each published record is such that every combination of values of quasi-identifiers can be matched to at least k respondents. This idea is easy to understand and many types of k-anonymization algorithms have been proposed: The Incognito algorithm [11] generalizes the attributes using taxonomy trees and the Mondrian algorithm [12] averages or replaces the original data for representative values and achieves k-anonymization. In this paper, we use a k-anonymization algorithm based on clustering and denote $A_k(D)$ as k-anonymization for dataset D. The algorithm finds close records and consists of clusters, such that each partition contains at least k records. For details of the algorithm, see [13].

2.1.2 Noise Addition

Noise addition works by adding or multiplying stochastic or randomized numbers to confidential data [14]. The idea is simple and is also well known as an anonymization technique. The first work on noise addition was proposed by Kim [15] and the idea was to add noise ϵ with distribution $\epsilon \sim N(0, \sigma^2)$ to original data. Additive noise is uncorrelated noise and preserves the mean and covariance of the original data but the correlation coefficients and variances are not sustained. Another variation of additive noise is correlated additive noise that keeps the mean and allows the correlation coefficients in the original data to be sustained [16]. Differential privacy is a state-of-the-art privacy model that is based on the statistical distance between two database tables differing by at most one record. The basic idea is that, regardless of background knowledge, an adversary with access to the dataset draws the same conclusions, irrespective of whether a person's data is included in the dataset. Differential privacy is mainly studied in relation to perturbation methods in an interactive setting, although it is applicable to certain generalization methods.

In this paper, we use Laplace noise as a noise addition and add noise $\epsilon \sim Lap(0, 2\phi^2)$ to each attribute. We can apply any types of noise addition to evaluate the privacy risk. Compared with normal distribution, Laplace noise has small effect to many records and we can obtain better results in experiments when we use Laplace noise.

We denote $A_\phi(D)$ as noise addition for a dataset D.

2.2 Matrix Factorization

Matrix factorization is a fundamental task in data analysis and the technique is used in various scenes, such as text data mining, acoustic analysis, and product recommendation by collaborative filtering. We use a matrix factorization as an anonymization technique, so we present the overview of a matrix facorization in this section.

2.2.1 SGD Matrix Factorization

Let an unknown rank-r matrix be $M \in \mathbb{R}^{n \times m}$, of which we know set $\Omega \subset [n] \times [m]$ of elements. $P_\Omega(M) \in \mathbb{R}^{n \times m}$ is defined below:

$$P_\Omega(M) = \begin{cases} M_{ij} & \text{if } (i, j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The goal of a matrix factorization is to find two matrices $U \in \mathbb{R}^{r \times n}$ and $V \in \mathbb{R}^{r \times m}$ which approximate the original matrix $M_{ij} \approx X_{ij}$ s.t. $\forall M_{ij} \in \Omega(M)$ with lower dimensionality $r << min(n, m)$. Here, $X = U^{\top}V$.

This problem is defined to solve the following optimization problem:

$$\min_{u^*,v^*} \sum_{(i,j)\in P_\Omega(M)} (M_{ij} - u_i^{\mathsf{T}} v_j)^2 + \lambda(\|u_i\|^2 + \|v_j\|^2), \qquad (2)$$

where u_i is a user-factors vector and v_j is an item-factors vector. When u_i and v_j are variables, this function is not a convex set, so that the problem described above cannot be solved. The second term of Eq. (2) is for regularization and the regularization parameter λ prevents overfitting. In this paper, we consider the rank r is a parameter of matrix factorization as an anonymization method, but actually, λ can suppress the complexity of the model and we can consider that λ is also a parameter of matrix factorization. However, the effect of plural parameters is difficult to evaluate, so that we fix $\lambda = 0.5$, which is the typical value, and consider only rank r is the parameter of matrix factorization. Equation (2) is widely used and you can find details in some references (e.g. [17]). Some techniques are proposed to solve the problem and gradient descent [18], for example, is a fundamental technique to find a local minimum value. However, it needs to update the vectors iteratively to obtain an optimal solution and using gradient descent is computationally expensive, so that stochastic gradient descent (SGD) is widely used such as in KDDCup 2011 [19] and Netflix Prize [17].

There are some research to speed up SGD-based matrix factorization such as [20]–[23] and each algorithm updates the matrices in parallel or in a distributed manner.

In this paper, we apply simple SGD technique to optimize the formular (2) and denote *Update*(A) to update a matrix A using SGD technique.

2.3 Matrix Factorization for Time-Sequence Data

There are some studies using matrices for time-sequence data. Zheng et al. [3], [24] proposed to predict a user's interests in an unvisited location. They assumed users' GPS trajectory as a user-location matrix and each value of the matrix means the number of visits of a user to a location. The matrix is very sparse because each user visits only a handful of locations, so a collaborative filtering model is applied to the prediction. Zheng et al. [4] built a location-activity matrix, M, which has missing values. M is decomposed into the two low-rank matrices, U and V. The missing values can be filled by $X = UV^{\mathsf{T}} \simeq M$ and the locations can be recommended when some activities are given. Chawla et al. [5] constructed a graph from trajectories of taxis and transformed the graph into matrices. The authors of [1] proposed a method of identifying the traffic flows that cause an anomaly between two regions.

3. Definitions

3.1 Privacy Definition

We define two types of attack models for time-sequence datasets. The first one, re-identification attack, is a general attack model and an attacker has information on the original dataset M and tries to re-identify it in an anonymized dataset $A(M)$. This model assumes that an attacker has maximum information about the original dataset. This model is same as that of k-anonymization where even if an attacker has an original dataset, the probability of re-identification of a k-anonymized dataset is $1/k$.

Definition 1 *Re-identification attack*: Let an attacker have matrix $M_{t_1} \in \mathbb{R}^{n\times m}$ and an anonymized matrix $A(M_{t_1}) \in \mathbb{R}^{n\times m}$. M_{t_1} represents a time-sequence data, which is observed during t_1, and n is the number of records and m is the number of items. A re-identification attack against a record r_i succeeds if record $r_i \in M_{t_1}$ is linked to record $r'_j \in A(M_{t_1})$, where r'_j is the anonymized r_i or belongs to the cluster which includes the anonymized r_i.

The linkage attack, which is the attack of an authorized user, is that an attacker tries to obtain some information from the given datasets $A(M_{t_1})$ and $A(M_{t_2})$. $A(M_{t_1})$ and $A(M_{t_2})$ are assumed to be included the same users but the primary keys are different. An attacker in this model has only anonymized datasets, so that an authorized user is assumed to be an attacker in this model. There are few studies concerning this problem and we evaluate the risk using actual datasets in this paper.

Definition 2 *Linkage attack*: Let an attacker have two anonymized matrix $A(M_{t_1}) \in \mathbb{R}^{n\times m}$ and $A(M_{t_2}) \in \mathbb{R}^{n\times m}$. M_T represents a time-sequence data, which is observed during T, n is the number of records, m is the number of items and both M_{t_1} and M_{t_2} include the same users and items. A linkage attack against a record r_i succeeds if record $r'_i \in A(M_{t_1})$ is linked to record $r''_j \in A(M_{t_2})$, where r'_i and r''_j represent the same user or r''_j belongs to the cluster which includes the same user of r'_i.

We next define privacy metric as follows:

Definition 3 *Privacy metric*: Let n be the total number of users of a dataset M and n' be the number of users which are attacked succesfully. The privacy risk of M is defined $\frac{n'}{n}$.

Example: We consider the attacks to be the same as those to solve an assignment problem. An assignment problem is to find the task assignment properly when there are n users and tasks and the Hungarian algorithm [25] solves the assignment problem in such a way that the entire cost is minimal. We apply the algorithm as re-identification and linkage attacks and consider when an attacker assigns the same user, the attack succeeds. When a dataset is k-anonymized, there are the same records at least $k-1$. Hence, when a record is assigned to the cluster that the correct record belongs to, we regard the record as being assigned correctly even if the assigned record is not actually correct. Furthermore, we define the privacy metric as the result obtained by multiplying the probability and $1/k$ because the probability means the ratio of correct assignment of clusters.

Figure 1 shows an example of risk evaluation. The dataset on the left is the original dataset and that on the right is the anonymized dataset. The arrows indicate the assign-

User	Data
1	1.0
2	1.5
3	1.5
4	2.5
5	3.5
6	5.0
7	6.0

User	Data
1	1.25
2	1.25
3	2.5
4	2.5
5	2.5
6	5.5
7	5.5

Fig. 1 Example of a risk evaluation.

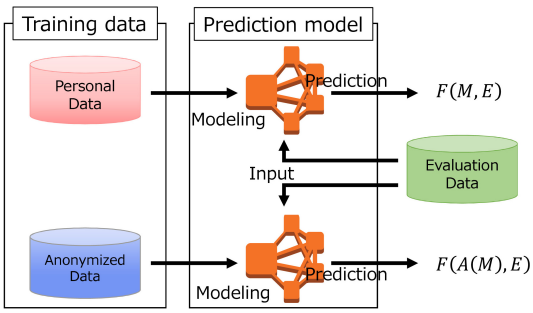

Fig. 2 Overview of utility evaluation.

ment result. User 2 of the original dataset, for instance, is assigned to user 3 of the anonymized dataset, so the attack for user 2 fails. When noise addition is used as the anonymization method, users 2, 3, 4, and 5 are assigned to the wrong users and the privacy risk is 3/7. On the other hand, when k-anonymization is used, in this case $k = 2$, users 4 and 5 are assigned to the wrong users (blue arrows) but assigned to the clusters that are the same as the correct users. Therefore, we consider the attack for users 4 and 5 to be successful. The failed attacks are only for users 2 and 3 (red arrows) and the privacy risk is $5/7 \times 1/2 = 5/14$.

3.2 Utility Definition

We define the utility metric here. In previous research, most utility metrics are based on either the distance between the original dataset and the anonymized dataset, or the amount of information loss [13], [26]. However, the utility depends on the situation (i.e., context, use-case) and these metrics do not necessarily match the actual utility. Therefore, we consider a use-case scenario and present a utility definition that matches the scenario. Specifically, we consider a use-case in which an anonymized dataset is used as training data for a machine learning algorithm. In the case of a web access log dataset, for example, a client, who is a developer of an anti-virus software, may generate a machine learning model from an anonymized dataset and predict whether their user will access a phishing website.

Definition 4 *Utility metric*: Let $F(M, E)$ be the F-measure of a machine learning model, where the training data is M and the test data is E. The utility metric is defined as follows:

$$Uti(A(M)) = \frac{F(A(M), E)}{F(M, E)}, \tag{3}$$

where $A(M)$ is an anonymized M.

We consider an actual use-case and generate machine learning model. When we consider the risk of documents, it is diffucult to evaluate the utility because documents have been anonymized subjectively so far. Therefore, some researchers such as [27] use F-measure to evaluate the utility. On the other hand, we apply F-measure assuming actual usage. Figure 2 gives an overview of the utility evaluation. We first generate two machine learning models; one is from an original dataset and the other is from its anonymized dataset. An item is randomly chosen as a objective variable and the remainders are explanation variables. And then, we use these models and predict an attribute of each record of an evaluation dataset that has the same attributes as the original dataset. This operation is performed several times while an objective variable is changed. The utility is defined as the average of the ratio of the F-measure of a model of the anonymized dataset to that of a model of the corresponding original dataset. In this paper, we apply logistic regression as the machine learning algorithm and predict for fifty attributes.

4. Anonymization Using Matrix Factorization

We consider matrix factorization as an anonymization method and rank r contributes to the accuracy of the matrix approximation. Moreover, we combine anonymization methods as same as some previous studies [28], [29]. Specifically, we propose to combine matrix factorization with another anonymization method *ano*, such as k-anonymization and noise addition. We denote p as a parameter of the anonymization method and p is k or ϕ in this paper. Basis matrix U and weighting matrix V can be assumed as the characteristics of rows and columns, respectively, and U is a characteristic matrix of users in our dataset. Therefore, we propose to anonymize U and maintain V so that the characteristics of domain is preserved. In our algorithm, we first divide the dataset M into U and V, and anonymize U. After that, we optimize V once and recombine it with the anonymized U. The algorithm is described as follows.

We indicate that $A_r(D)$ applies matrix factorization to matrix D and that $A_{(ano,r)}(D)$ combines matrix factorization and the anonymization method *ano* as follows.

$$A_{(ano,r)}(D) = (A_{(ano)}(U))^{\mathsf{T}} V, \text{ where } U \in \mathbb{R}^{r \times n}, \ V \in \mathbb{R}^{r \times m}. \tag{4}$$

5. Experiment

5.1 Dataset

We use an actual web access log dataset as a time-sequence dataset. The dataset consists of an ID, a time-stamp, and the

Algorithm 1 (M, r, I, ano, p): Anonymization using Matrix Factorization

Require: Original dataset M, rank r, anonymization function and the parameter (ano, p), and the number of iteration I.

1: $t = 0$
2: Construct $U_t \in [0,1]^{n \times r}$ and $V_t \in [0,1]^{m \times r}$ randomly
3: **while** $t < I$ **do**
4: $\quad U_{t+1} = Update(U_t)$
5: $\quad V_{t+1} = Update(V_t)$
6: $\quad t = t + 1$
7: **end while**
8: $U'_{t+1} = A_{(ano)}(U_{t+1})$
9: **return** $X = U'^{\mathsf{T}}_{t+1} V_{t+1}$

Table 1 Dataset format.

ID (= i)	Date	URL (= j)
x_{t_1} (= 1)	2016-12-01 16:13:48	www.google.com (= 1)
y_{t_1} (= 2)	2016-12-01 16:15:14	mail.google.com (= 2)
x_{t_1}	2016-12-01 16:17:13	www.youtube.com (= 3)
z_{t_1} (= 3)	2016-12-01 16:19:01	www.facebook.com (= 4)
x_{t_2} (= 1)	2016-12-01 16:21:15	www.youtube.com
x_{t_2}	2016-12-01 16:22:42	www.google.com
z_{t_2} (= 3)	2016-12-01 16:25:01	www.youtube.com

access domain as shown in Table 1. We convert the dataset into a matrix as follows.

$$M_T = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{5}$$

Here, T is the observation time.

We denote $r_{ij} = 1$ if a user whose ID is i accesses domain j during time T, and otherwise $r_{ij} = 0$. For example, we denote the datasets in Table 1 as follows.

$$M_{t_1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

$$M_{t_2} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{7}$$

Here, t_1 is the 10-minutes span between 2016-12-01 16:10:00 and 2016-12-01 16:19:59, and t_2 is the similar 10-minutes span between 2016-12-01 16:20:00 and 2016-12-01 16:29:59. The IDs are different between t_1 and t_2 but x_{t_1} and x_{t_2}, and z_{t_1} and z_{t_2} represent the same users.

In the following experiments, we chose 200 users and 1,000 domains randomly from an actual web access log and let the pseudonymous ID be changed at every designated time T.

5.2 The Privacy Risk against the Linkage Attack

First, we evaluate whether a linkage attack is possible or not. We set the observation time t_1 as 2, 4, and 8 hours from 16:00 on a weekday and observation time t_2 as the same time on another weekday. The probability of a linkage attack between M_{t_1} and M_{t_2} is shown in Table 2.

Table 2 Linkage attack against non-anonymized dataset.

Observation Time	Risk against Linkage attack
$2h$	0.51
$4h$	0.64
$8h$	0.80

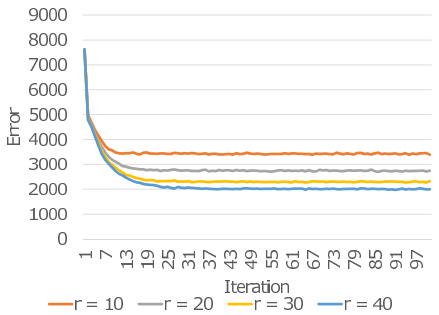

Fig. 3 The relationship between the error and the number of iteration.

The matrix only includes the information on whether a domain has been accessed or not, and even if the observation time is 2 hours, the linkage attack probability, i.e., risk, is very high (over 50%). Moreover, the risk rises as the observation time increases because when the observation time increases, the tendency of a user becomes remarkable. The result shows that the pattern of web access for people bear consistent characteristics. Hence, we need to care not only of the re-identification attack but also of the linkage attack so as to avoid privacy leakages.

5.3 Effects of Matrix Factorization Itself

Observation time t_1 and t_2 are fixed as 8 hours from 16:00 hours on a weekday in the following experiments. The inputs of matrix factorization are original dataset $M \in \mathbb{R}^{200 \times 1000}$, number of iterations I, and rank r. Furthermore, λ and γ are the hyper parameters. When $\gamma = 0.05$, and $\lambda = 0.5$, the relationship between the error, namely $\sum_{ij} |(M_{t_1})_{ij} - (X_{t_1})_{ij}|$, and the iteration number is in Fig. 3. The figure shows the error is almost fixed when the number of iteration is over 25. Hence, we fix $I = 100$, which is enough to converge. Rank r can be treated as the parameter of anonymization by matrix factorization because the accuracy of dataset $X = UV^{\mathsf{T}}$ depends on rank r, so that r is the parameter of our algorithm and we set $r = 10, 20, 30, 40$. We set larger values in the experiments in [7] but the results of the case $r > 40$ are saturated. The probabilities of re-identification and linkage attack are shown in Table 3.

The results show that matrix factorization itself does not have much effect on re-identification attacks. Note that matrix factorization can preserve the relative positional relationship among the records, so that the privacy risk of the re-identification attack does not decrease so much by a matching algorithm. When the rank is small enough, $r = 10$, the positional relationship is broken and the privacy risk is low-

Table 3 Attacks against matrix factorization.

Rank	Risk against Re-identification attack	Risk against Linkage attack
10	0.98	0.31
20	1.00	0.45
30	1.00	0.54
40	1.00	0.58

ered.

On the other hand, compared with the re-identification attack in Table 2, the linkage attack probability between $A_r(M_{t_1})$ and $A_r(M_{t_2})$ is better. This is because the relationship between the records of M_{t_1} and M_{t_2} is weaker than that of between M_{t_1} and $A_r(M_{t_1})$. In our experiment, the dataset of the observation time is 8 hours and $r = 30$ has almost the same privacy level as that of the observation time is 2 hours.

5.4 Risk Evaluation

We evaluate our anonymization method, algorithm 1, in the following experiments. In the following experiments, we apply [13] as k-anonymization and Laplace noise as a noise addition. When noise addition is applied, noise $\epsilon \sim Lap(0, 2\phi^2)$ is added to each element and the parameter is ϕ.

1. Evaluate the privacy risk of re-identification attack between $A_k(M_{t_1})$ and M_{t_1}, and linkage attack between $A_k(M_{t_1})$ and $A_k(M_{t_2})$.
2. Evaluate the privacy risk of re-identification attack between $A_\phi(M_{t_1})$ and M_{t_1}, and linkage attack between $A_\phi(M_{t_1})$ and $A_\phi(M_{t_2})$.
3. Evaluate the privacy risk of re-identification attack between $A_k(U_{t_1})^\mathsf{T}V$ and M_{t_1}, and linkage attack between $A_k(U_{t_1})^\mathsf{T}V$ and $A_k(U_{t_2})^\mathsf{T}V$.
4. Evaluate the privacy risk of re-identification attack between $A_\phi(U_{t_1})^\mathsf{T}V$ and M_{t_1}, and linkage attack between $A_\phi(U_{t_1})^\mathsf{T}V$ and $A_\phi(U_{t_2})^\mathsf{T}V$.

The evaluations of the re-identification attack in experiments 1 and 2 are almost the same as those conducted in many previous research. The difference is the privacy metric (see Sect. 3.1) and these results are used for comparison with experiments 3 and 4, which are the evaluations of our algorithm. There are few studies on linkage attacks, and evaluations of the attack are one of our contributions.

The evaluation of re-identification attack in experiment 1 (Fig. 4) is simple and the result is almost the same as k-anonymization. However, our privacy metric is a little different from that for k-anonymity, so the result is also a little different from $1/k$. The result of the linkage attack also shows that k-anonymization can greatly improve the privacy of linkage attack and 2-anonymization can reduce the privacy lisk by 77% (0.8 → 0.185).

The evaluations in experiment 2 are shown in Table 5. The privacy of the re-identification attack is improved from $\phi \geq 0.9$ and when ϕ is large, for example $\phi = 1.5$, the score looks good. However, almost the half records are changed more than 1 by the adding noise and the each original value of M is 0 or 1, namely, $M_{ij} \in \{0, 1\}$, so that the noise is too large to preserve the utility. Therefore, we conclude that simple noise addition is not good, in terms of utility preservation, as an anonymization method. On the other hand, we obtain an interesting result for linkage attack. The privacy for linkage attack is improved even if the noise is very small and we can say that the adding even small noise is an effective countermeasure against the linkage attack.

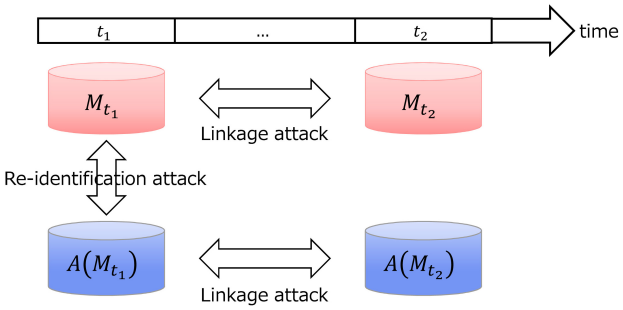

Fig. 4 Overview of experiment.

Table 4 Experiment 1: The privacy risk of a k-anonymized data.

k	Re-identification Attack	Linkage Attack
2	0.500	0.185
4	0.250	0.050
6	0.167	0.038
8	0.125	0.027
10	0.098	0.023

Table 5 Experiment 2: The privacy risk of a noise added data.

ϕ	Re-identification Attack	Linkage Attack
0.3	1.00	0.33
0.6	1.00	0.10
0.9	0.95	0.01
1.2	0.81	0.03
1.5	0.62	0.00

Table 6 Experiment 3: The privacy risk of a data applied with Algorithm 1 (SGD + k-anonymization) for Re-identification attack.

k	$r = 10$	$r = 20$	$r = 30$	$r = 40$
2	0.44	0.50	0.50	0.50
4	0.21	0.24	0.25	0.25
6	0.12	0.14	0.15	0.16
8	0.10	0.11	0.11	0.12
10	0.08	0.08	0.08	0.08

In experiment 3, we evaluate the effect of our proposed algorithm, which is a combined matrix factorization and k-anonymization. Table 6 is the result of the re-identification attack. In the experiment, we cannot find the effect of the matrix factorization a lot but the privacy improves slightly as r increases. This is because, k-anonymization has a large effect on the re-identification risk and the effect of the matrix factorization does not appear.

The results of linkage attack in experiment 3 are shown in Table 7. In the experiment, we cannot obtain new knowledge about the effect of the matrix factorization. When the datasets, which are observed at different time periods, are

Table 7 Experiment 3: The privacy risk of a data applied with Algorithm 1 (SGD + k-anonymization) for Linkage attack.

k	$r = 10$	$r = 20$	$r = 30$	$r = 40$
2	0.11	0.15	0.15	0.15
4	0.05	0.07	0.08	0.07
6	0.04	0.03	0.03	0.04
8	0.03	0.03	0.03	0.03
10	0.02	0.02	0.02	0.02

Table 8 Experiment 4: The privacy risk of a data applied with Algorithm 1 (SGD + noise addition) for Re-identification attack.

ϕ	$r = 10$	$r = 20$	$r = 30$	$r = 40$
0.05	0.75	0.95	0.97	1.00
0.10	0.42	0.72	0.85	0.86
0.15	0.25	0.50	0.61	0.70
0.20	0.18	0.28	0.40	0.49

Table 9 Experiment 4: The privacy risk of a data applied with Algorithm 1 (SGD + noise addition) for Linkage attack.

ϕ	$r = 10$	$r = 20$	$r = 30$	$r = 40$
0.05	0.21	0.34	0.34	0.50
0.10	0.12	0.15	0.14	0.20
0.15	0.07	0.11	0.09	0.10
0.20	0.03	0.03	0.03	0.02

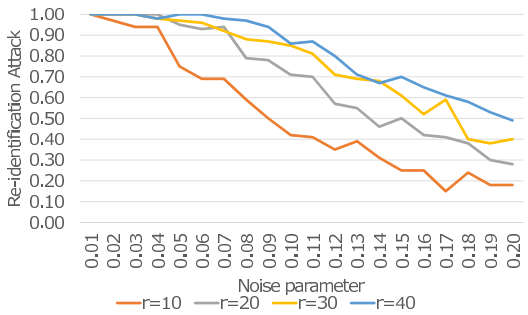

Fig. 5 The re-identification risk of combination of matrix factorization and noise addition.

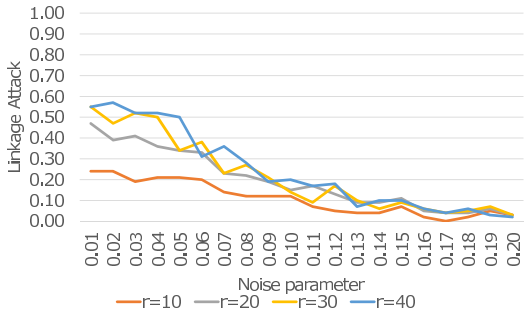

Fig. 6 The linkage risk of combination of matrix factorization and noise addition.

sufficiently anonymized by k-anonymization, there is no relation among the same users of each dataset and only outliers can be linked.

In experiment 4, we evaluate the impact of our method, which is a combination of matrix factorization and noise addition. The evaluation results of the re-identification attack are shown in Table 8. The noise is added to U, which is the user's characteristics, and then, U^{T} is multiplied with V. Therefore, we cannot compare the results with those of experiment 2 simply, but the impact of the matrix factorization is high. This result shows that using matrix factorization can help to construct anonymized datasets flexibly from the viewpoint of privacy. For example, the privacy risk of $A_{(\phi=0.15,r=20)}(M_{t_1})$ and $A_{(\phi=0.20,r=40)}(M_{t_1})$ is almost the same as that of $A_{(k=2)}(M_{t_1})$ and $A_{(\phi=1.5)}(M_{t_1})$.

The results of the linkage attack in experiment 4 are described in Table 9. The tendency is the same as that of re-identification attack and the matrix factorization is compatible with the noise addition. We present the details of the results of the re-identification attack and the linkage attack in Figs. 5, 6.

5.5 Utility Evaluation

We next evaluate the utility of anonymized datasets. We evaluate the utility of datasets applying a machine learning algorithm. A logistic regression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) is applied in the following experiment and the parameters are default setting. One of utilizations of an access log dataset is to prevision a malignant site and inform the web browser's users. Therefore, we use a machine learning algorithm and predict whether each user will access a malignant site or not. We generate learning models using the original (non-anonymized) dataset and the anonymized datasets and input the test dataset into these models. The utility score is defined in Definition 4 and the F-measure of the model of the original dataset was 0.763. Each result of the evaluation is shown in Tables 10, 11, 12, and 13.

1. Evaluate the utlity of $A_{(k)}(M_{t_1})$ for $k = 2, 4, 6, 8, 10$.
2. Evaluate the utlity of $A_{(\phi)}(M_{t_1})$ for $\phi = 0.3, 0.6, 0.9, 1.2, 1.5$.
3. Evaluate the utlity of $A_{(k=2,r)}(M_{t_1})$ for $r = 10, 20, 30, 40$.
4. Evaluate the utlity of $A_{(\phi,r)}(M_{t_1})$ for $\phi = 0.1, 0.15$ and $r = 10, 20, 30, 40$.

The actual dataset we use is $M_{ij} \in \{0, 1\}$ and the matrix is sparse. Moreover, some people have the same tendency in the dataset we use, so that k-anonymization is effective in our experiment. However, when the dataset is more complex such as image data, which can not be expressed only by 0 or 1, the utility of k-anonymization will decrease.

The results of the experiment 2 shows that the utility of the dataset decreases as the noise increases. As denoted in the risk evaluation section, each element of the original dataset is 0 or 1 and the utility gets worse drastically when the noise parameter is large such as $\phi = 1.5$.

When k-anonymization and the matrix factorization is combined, the effect of the matrix factorization is small as well as the case of the privacy risk. In this experiment, the effect of k-anonymization is large and the effect of the matrix factorization is relatively small.

Table 10 Utility Evaluation 1: The utility of k-anonymized data.

Dataset D	Precision	Recall	F measure	$Uti(D)$
$A_{(k=2)}(M_{t_1})$	0.780	0.720	0.749	0.981
$A_{(k=4)}(M_{t_1})$	0.741	0.688	0.714	0.936
$A_{(k=6)}(M_{t_1})$	0.755	0.691	0.721	0.946
$A_{(k=8)}(M_{t_1})$	0.737	0.659	0.696	0.913
$A_{(k=10)}(M_{t_1})$	0.748	0.677	0.711	0.932

Table 11 Utility Evaluation 2: The utility of noise added data.

Dataset D	Precision	Recall	F measure	$Uti(D)$
$A_{(\phi=0.3)}(M_{t_1})$	0.780	0.664	0.717	0.941
$A_{(\phi=0.6)}(M_{t_1})$	0.738	0.610	0.668	0.876
$A_{(\phi=0.9)}(M_{t_1})$	0.719	0.541	0.618	0.810
$A_{(\phi=1.2)}(M_{t_1})$	0.652	0.507	0.571	0.748
$A_{(\phi=1.5)}(M_{t_1})$	0.625	0.520	0.567	0.744

Table 12 Utility Evaluation 3: The utility of data applied with Algorithm 1 (SGD + k-anonymization)

Dataset D	Precision	Recall	F measure	$Uti(D)$
$A_{(k=2,r=10)}(M_{t_1})$	0.686	0.735	0.710	0.930
$A_{(k=2,r=20)}(M_{t_1})$	0.699	0.767	0.731	0.959
$A_{(k=2,r=30)}(M_{t_1})$	0.695	0.773	0.732	0.960
$A_{(k=2,r=40)}(M_{t_1})$	0.712	0.786	0.747	0.980

Table 13 Utility Evaluation 4: The utility of data applied with Algorithm 1 (SGD + noise addition)

Dataset D	Precision	Recall	F measure	$Uti(D)$
$A_{(\phi=0.10,r=10)}(M_{t_1})$	0.742	0.650	0.693	0.909
$A_{(\phi=0.10,r=20)}(M_{t_1})$	0.752	0.688	0.719	0.943
$A_{(\phi=0.10,r=30)}(M_{t_1})$	0.736	0.703	0.719	0.943
$A_{(\phi=0.10,r=40)}(M_{t_1})$	0.737	0.735	0.736	0.965

Table 14 Utility Evaluation 5: The utility of data applied with Algorithm 1 (SGD + noise addition)

Dataset D	Precision	Recall	F measure	$Uti(D)$
$A_{(\phi=0.15,r=10)}(M_{t_1})$	0.718	0.614	0.662	0.868
$A_{(\phi=0.15,r=20)}(M_{t_1})$	0.748	0.655	0.698	0.915
$A_{(\phi=0.15,r=30)}(M_{t_1})$	0.704	0.680	0.692	0.907
$A_{(\phi=0.15,r=40)}(M_{t_1})$	0.716	0.711	0.713	0.935

The evaluaton results of the combination of the noise addition and the matrix factorization present a good parformance (Tables 13 and 14). A dataset generated by combining the matrix factorization and noise addition has more utility than a dataset generated by noise addition when each dataset has the same privacy level.

In the experiments, we can say that our proposal algorithm has at least three strengths. Firstly, the proposed algorithm can control the privacy risk using the parameter r flexibly. For example, Fig. 5 shows that the privacy improves as the rank r becomes smaller. Secondly, matrix factorization itself is efficient when we consider a linkage attack model. The relationship between the records of M_{t_1} and M_{t_2} is weak, so that the privacy for linkage attack can be improved easily by using matrix factorization. Finally, the proposed algorithm improves the privacy of a dataset while maintaining the utility of the dataset especially when the noise addition and the matrix factorization is combined. For example, the privacy risk and the utility of $A_{(\phi=1.5)}(M_{t_1})$ are 0.62 and 0.744. On the other hand, those of $A_{(\phi=0.15,r=30)}$ are 0.61 and 0.907. This means our proposal algorithm can improve the utility maintaining the privacy of the dataset.

6. Conclusion

In this paper, we proposed an anonymization method using matrix factorization. Moreover, we conducted some experiments and showed that an anonymization method combining a matrix factorization and noise addition can maintain higher utility than only noise addition. Furthermore, we consider the risk of the linkage between the same records that are pseudonymized. The experimental results show that the linkage risk remains if the anonymization is insufficient, but the privacy can be improved by noise addition for instance, even if it is very small.

Acknowledgments

A part of this work was partly supported by CREST (JPMJCR1404) at Japan Science and Technology Agency, enPiT (Education Network for Practical Information Technologies) at MEXT, Innovation Platform for Society 5.0 at MEXT, and WarpDrive: Web-based Attack Response with Practical and Deployable Research InitiatiVE at National Institute of Information and Communications Technology (NICT).

References

[1] Y. Zheng, "Trajectory data mining: An overview," ACM Trans. Intelligent Systems and Technology, vol.6, no.3, Article No.29, 2015.
[2] J. Krumm, "Inference attacks on location tracks," International Conference on Pervasive Computing, Pervasive 2007, Lecture Notes in Computer Science, vol.4480, pp.127–143, Springer, Berlin, Heidelberg, 2007.
[3] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," Proc. 18th International Conference on World Wide Web, pp.791–800, ACM, 2009.
[4] V.W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with GPS history data," Proc. 19th International Conference on World Wide Web, pp.1029–1038, ACM, 2010.
[5] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," 2012 IEEE 12th International Conference on Data Mining (ICDM), pp.141–150, IEEE, 2012.
[6] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," 2008 IEEE 24th International Conference on Data Engineering, pp.376–385, 2008.
[7] T. Mimoto, S. Kiyomoto, S. Hidano, A. Basu, and A. Miyaji, "The possibility of matrix decomposition as anonymization and evaluation for time-sequence data," 2018 16th Annual Conference on Privacy, Security and Trust, pp.1–7, IEEE, 2018.
[8] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," Proc. PODS 1998, p.188, 1998.
[9] P. Samarati, "Protecting respondents identities in microdata release," IEEE Trans. Knowl. Data Eng., vol.13, no.6, pp.1010–1027, 2001.
[10] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," Int. J. Uncertainty, Fuzziness and

Knowledge-Based Systems, vol.10, no.5, pp.571–588, 2002.
[11] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain *k*-anonymity," Proc. SIGMOD 2005, pp.49–60, 2005.
[12] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," Proc. 22nd International Conference on Data Engineering (ICDE '06), pp.25–35, IEEE, 2006.
[13] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," International Conference on Database Systems for Advanced Applications, DASFAA 2007, Lecture Notes in Computer Science, vol.4443, pp.188–200, Springer, Berlin, Heidelberg, 2007.
[14] K. Mivule, "Utilizing noise addition for data privacy, an overview," arXiv preprint arXiv:1309.3958, 2013.
[15] J.J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation," Proc. Section on Survey Research Methods, pp.303–308, American Statistical Association, 1986.
[16] T. Yu and S. Jajodia, Secure Data Management in Decentralized Systems, Advances in Information Security, vol.33, Springer Science & Business Media, 2007.
[17] R.M. Bell and Y. Koren, "Lessons from the Netflix prize challenge," SIGKDD Expl. Newsl., vol.9, no.2, pp.75–79, 2007.
[18] J. Nocedal and S.J. Wright, Numerical Optimization, Springer Series in Operations Research and Financial Engineering. Springer Science & Business Media, 2006.
[19] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer, "The yahoo! music dataset and kdd-cup'11," Proc. 2011 International Conference on KDD Cup 2011-Volume 18, pp.3–18, JMLR.org, 2011.
[20] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," Advances in Neural Information Processing Systems, pp.693–701, 2011.
[21] R. Gemulla, E. Nijkamp, P.J. Haas, and Y. Sismanis, "Large-scale matrix factorization with distributed stochastic gradient descent," Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.69–77, ACM, 2011.
[22] Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin, "A fast parallel SGD for matrix factorization in shared memory systems," Proc. 7th ACM Conference on Recommender Systems, pp.249–256, ACM, 2013.
[23] J. Oh, W.-S. Han, H. Yu, and X. Jiang, "Fast and robust parallel SGD matrix factorization," Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.865–874, ACM, 2015.
[24] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," ACM Trans. Web, vol.5, no.1, Article No.5, 2011.
[25] H.W. Kuhn, "The Hungarian method for the assignment problem," Naval Research Logistics, vol.2, no.1-2, pp.83–97, 1955.
[26] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, "Utility-based anonymization for privacy preservation with less information loss," SIGKDD Explor. Newsl., vol.8, no.2, pp.21–30, 2006.
[27] D. Sánchez and M. Batet, "C-sanitized: A privacy model for document redaction and sanitization," Journal of the Association for Information Science and Technology, vol.67, no.1, pp.148–163, 2016.
[28] T. Mimoto, S. Kiyomoto, K. Tanaka, and A. Miyaji, "(p, n)-identifiability: Anonymity under practical adversaries," 2017 IEEE Trustcom/BigDataSE/ICESS, pp.996–1003, 2017.
[29] M. Freiman, J. Lucero, L. Singh, J. You, M. DePersio, and L. Zayatz, "The microdata analysis system at the U.S. census bureau," SORT Special issue: Privacy in statistical databases, pp.77–98, 2011.

Tomoaki Mimoto received his bachelors degree in engineering from Osaka university, Japan, in 2012, and received his master degree (Outstanding Performance Award) in information science from Japan Advanced Institute of Science and Technology in 2014. He joined KDDI in 2014, and has been with the KDDI research, Inc. since 2015, and is currently an associate research engineer in the Information Security Group.

Seira Hidano received his M.E. and Ph.D. degrees in computer science and engineering from Waseda University, Japan, in 2009 and 2012, respectively. In 2010, he was a JSPS research fellow. In 2011 and 2012, he was a research assistant at Waseda University. In 2013, he joined KDDI. He is currently a research engineer of the Information Security Lab. in KDDI Research, Inc. His research interest includes biometric authentication, information theoretic security, and privacy preservation.

Shinsaku Kiyomoto received his B.E. in engineering sciences and his M.E. in Material Science from Tsukuba University, Japan, in 1998 and 2000, respectively. He joined KDD (now KDDI) and has been engaged in research on stream ciphers, cryptographic protocols, and mobile security. He is currently a senior manager at the Information Security Laboratory of KDDI R&D Laboratories Inc. He was a visiting researcher of the Information Security Group, Royal Holloway University of London from 2008 to 2009. He received his doctorate in engineering from Kyushu University in 2006. He received the IEICE Young Engineer Award in 2004, Distinguished Contributions Awards in 2011, and Achievement Award in 2016. He is a member of IEICE and JPS.

Atsuko Miyaji received the B.Sc., the M.Sc., and the Dr.Sci. degrees in mathematics from Osaka University, Osaka, Japan in 1988, 1990, and 1997 respectively. She joined Panasonic Co., LTD. from 1990 to 1998 and engaged in research and development for secure communication. She was an associate professor at the Japan Advanced Institute of Science and Technology (JAIST) in 1998. She joined the computer science department of the University of California, Davis from 2002 to 2003. She has been a professor at Japan Advanced Institute of Science and Technology (JAIST) since 2007 and the director of Library of JAIST from 2008 to 2012. She has been a professor at Graduate School of Engineering, Osaka University since 2015. Her research interests include the application of number theory into cryptography and information security. She received Young Paper Award of SCIS '93 in 1993, Notable Invention Award of the Science and Technology Agency in 1997, the IPSJ Sakai Special Researcher Award in 2002, the Standardization Contribution Award in 2003, Engineering Sciences Society: Certificate of Appreciation in 2005, the AWARD for the contribution to CULTURE of SECURITY in 2007, IPSJ/ITSCJ Project Editor Award in 2007, 2008, 2009, 2010, 2012, 2016, and the Director-General of Industrial Science and Technology Policy and Environment Bureau Award in 2007, Editorial Committee of Engineering Sciences Society: Certificate of Appreciation in 2007, DoCoMo Mobile Science Awards in 2008, Advanced Data Mining and Applications (ADMA 2010) Best Paper Award, The chief of air staff: Letter of Appreciation Award, Engineering Sciences Society: Contribution Award in 2012, Prizes for Science and Technology, The Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, International Conference on Applications and Technologies in Information Security (ATIS 2016) Best Paper Award, and The 16th IEEE Trustocm 2017 Best Paper Award, and IEICE milestone certification in 2017. She is a member of the International Association for Cryptologic Research, the Information Processing Society of Japan, and the Mathematical Society of Japan.